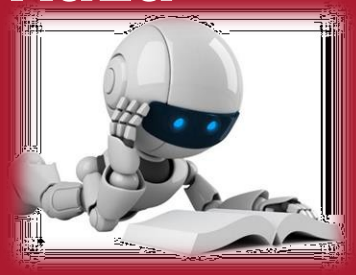
Evaluation of Classifiers

The Confusion Matrix

Asfia Ali Raza



The Output of a (

#	Height (inches)	Weight (kgs)	B.P. Sys →	B.P. Dia	Heart disease →	
	𝑋				𝑦	𝑦(𝑋)
1	62	70	120	80	No	No
2	72	90	110	70	No	Yes
3	74	80	130	70	No	No
4	65	120	150	90	Yes	Yes
5	67	100	140	85	Yes	No
6	64	110	130	90	No	Yes
7	69	150	170	100	Yes	Yes
8	75	127	160	95	Yes	No
9	66	66	135	90	Yes	Yes

- 𝑦: Gold standards / Gold labels / Ground truth
- 𝑦(𝑋): Predicted labels

The Output of a

#	Height (inches)	Weight (kgs)	B.P. Sys →	B.P. Dia	Heart disease →	
	𝒙				𝒚	𝒉(𝒙)
1	62	70	120	80	**No**	No
2	72	90	110	70	**No**	Yes
3	74	80	130	70	**No**	No
4	65	120	150	90	**Yes**	Yes
5	67	100	140	85	**Yes**	No
6	64	110	130	90	**No**	Yes
7	69	150	170	100	**Yes**	Yes
8	75	127	160	95	**Yes**	No
9	66	66	135	90	**Yes**	Yes

Negative
Positive
Negative
Positive
Negative
Positive
Positive
Negative
Positive

- $\vec{x}$: Gold standards / Gold labels / Ground truth
- $h(\vec{x})$: Predicted labels
  - Positives and Negatives

The Output of a

#	Height (cm)	Weight (kg)	B.P. Sys	B.P. Dia	Heart disease	□		
1	62	70	120	80	No	No	True	Negative
2	72	90	110	70	No	Yes	False	Positive
3	74	80	130	70	No	No	True	Negative
4	65	120	150	90	Yes	Yes	True	Positive
5	67	100	140	85	Yes	No	False	Negative
6	64	110	130	90	No	Yes	False	Positive
7	69	150	170	100	Yes	Yes	True	Positive
8	75	127	160	95	Yes	No	False	Negative

- □: Gold standards / Gold labels / Ground truth
- □(□): Predicted labels
  - Positives and Negatives
  - True (matching the ground truth) or False (not matching the ground truth)

The Output of a

#	Height (cm)	Weight (kg)	B.P. Sys	B.P. Dia	Heart disease	𝑦̂
1	62	70	120	80	**No**	No
2	72	90	110	70	**No**	Yes
3	74	80	130	70	**No**	No
4	65	120	150	90	**Yes**	Yes
5	67	100	140	85	**Yes**	No
6	64	110	130	90	**No**	Yes
7	69	150	170	100	**Yes**	Yes
8	75	127	160	95	**Yes**	No

True Negative

False Positive Type-I Error

True Negative

True Positive

False Negative Type-II Error

False Positive Type-I Error

True Positive

False Negative Type-II Error

- 𝑦: Gold standards / Gold labels / Ground truth
- 𝑦̂(𝑥): Predicted labels
  - Positives and Negatives
  - True (matching the ground truth) or False (not matching the ground truth)

#	Height (inches)	Weight	B.P. Sys	B.P. Dia	Heart disease	
1	62	70	120	80	No	No
2	72	90	110	70	No	Yes
3	74	80	130	70	No	No
4	65	120	150	90	Yes	Yes
5	67	100	140	85	Yes	No
6	64	110	130	90	No	Yes
7	69	150	170	100	Yes	Yes
8	75	127	140	80	Yes	Yes
9	66	66	135	90	Yes	Yes

Type-II errors

gold standard labels

system output labels	gold positive	gold negative
system positive	**true positive**	**false positive**
system negative	**false negative**	**true negative**

True	Negative		
False	Positive	Type-I	Error
True	Negative		
True	Positive		
False	Negative	Type-II	Error
False	Positive	Type-I	Error
True	Positive		
False	Negative	Type-II	Error
True	Positive		

Evaluation of Classifiers

Performance Measures

Asim Ali
Raza

CS535/EE514 – Machine

Evaluation of Classifiers

Evaluating a classifier means measuring how well a machine learning model predicts class labels. Because accuracy alone can be misleading—especially with imbalanced datasets—we use several metrics to get a complete picture of performance.

Common evaluation metrics include:

- **Accuracy** – overall correctness
- **Precision** – correctness of positive predictions
- **Recall (Sensitivity)** – ability to find actual positives
- **F1-Score** – harmonic mean of precision and recall
- **Specificity** – ability to identify actual negatives
- **AUC/ROC** – overall ranking performance

These metrics are derived from the **Confusion Matri**

		Gold Labels	
		Gold Positive	**Gold Negative**
Predicted Labels	**Predicted Positive**	True Positives (tp)	False Positives (fp)
	Predicted	False Negatives (fn)	True Negatives (tn)

False Negative Rate = "Negative Likelihood Ratio"
True Negative Rate

		Gold Labels	
		Gold Positive	**Gold Negative**
Predicted Label	**Predicted Positive**	True Positives (tp)	False Positives (fp)
	Predicted	False Negatives (fn)	True Negatives (tn)

$$Accuracy = \frac{tp + tn}{tp + fp + fn + tn}$$

$$Negative\ Predictive\ Value = \frac{tn}{fn + tn}$$

- You are the CEO of a Pie Company and you need to know what people are saying about your pies on social media
- You build a system that detects tweets concerning your pies
  - the positive class is tweets about your pies
  - the negative class is all other tweets
- Imagine that we looked at a million tweets
  - 100 of them are discussing their love (or hatred) for our pie
  - the other 999,900, are tweets about other matters

- You created a classifier that stupidly classified every tweet as “not about pie”
  - 999,900 true negatives, 100 false negatives, 0 false positives, and 0 true positives, for an accuracy of 999,900/1,000,000 or 99.99%!

- **Accuracy is not a good metric when the goal is to discover a rare event, or at least not completely balanced in**

Interpretation

- **TP**: Model correctly predicted positive.
- **TN**: Model correctly predicted negative.
- **FP (Type I Error)**: Model predicted positive incorrectly.
- **FN (Type II Error)**: Model predicted negative incorrectly.

Example: Credit Card Fraud Detection
You are the head of security for a credit card company.
Your goal: **detect fraudulent transactions**.

Dataset
You analyze **1,000,000 transactions**:
- **1,000** are actually fraudulent (positive class)
- **999,000** are legitimate (negative class)

Fraud is rare, but extremely important to catch.

A Useless Classifier
Your data scientist builds a classifier that labels **every transaction as "legitimate."**

Predicted ↓ / Actual →	**Actual Fraud**	**Actual Legitimate**
Predicted Fraud	TP = 0	FP = 0
Predicted Legitimate	FN = 1,000	TN = 999,000

Interpretation (with swapped orientation)

Predicted Fraud (top row)

- **TP = 0:** The model never correctly identified a fraud case.
- **FP = 0:** It also never wrongfully labeled a legitimate transaction as fraud.

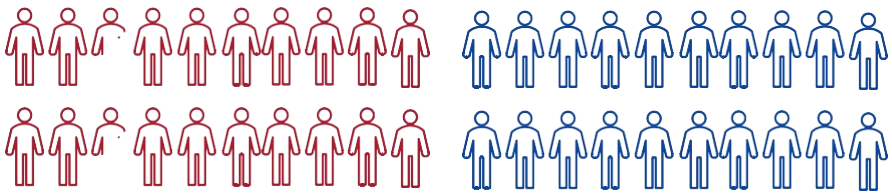
Predicted Legitimate (bottom row)

- **FN = 1,000:** All fraud cases were incorrectly labeled as legitimate.
- **TN = 999,000:** All legitimate transactions were correctly labeled.

Accuracy

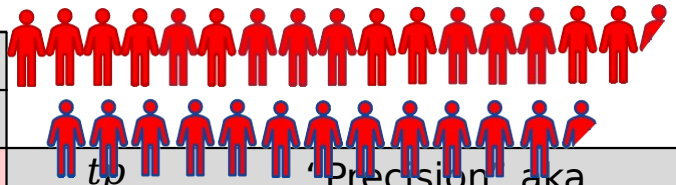
Looks great—but the classifier is **useless** because it misses **every single fraud case**.

		Gold Positive	Gold Negative		
		Gold Labels			
Predicted Labels	**Predicted Positive**	True Positives (tp)	False Positives (fp)	$\frac{tp}{tp + fp}$	"Precision" aka "Positive Predictive Value"
	Predicted Negative	False Negatives (fn)	True Negatives (tn)	$\frac{tn}{fn + tn}$	"Negative Predictive Value"
		$\frac{tp}{tp + fn}$	$\frac{tn}{fp + tn}$	$Accuracy = \frac{tp + tn}{tp + fp + tn + fn}$	

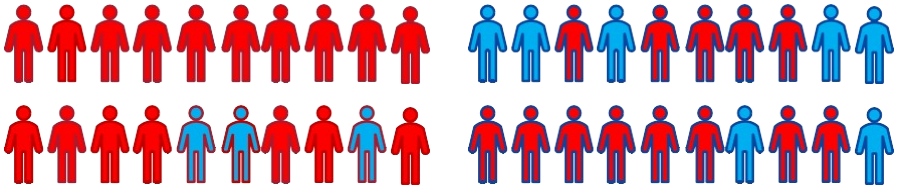


Likelihood Ratio" False Positive Rate

False Negative Rate = "Negative Likelihood Ratio"



		Gold Labels		
		Gold Positive	**Gold Negative**	
Predicted Labels	**Predicted Positive**	True Positives (tp)	False Positives (fp)	$\frac{tp}{tp + fp}$ "Precision" aka "Positive Predictive Value"
	Predicted Negative	False Negatives (fn)	True Negatives (tn)	$\frac{tn}{fn + tn}$ "Negative Predictive Value"
		$\frac{tp}{tp + fn}$	$\frac{tn}{fp + tn}$	$Accuracy = \frac{tp + tn}{tp + fp + tn + fn}$



True Positive Rate_ = "Positive Likelihood Ratio" False Positive Rate

Performance Metrics

Accuracy

$$= \frac{My\ Correct\ Answers}{All\ Questions} = \frac{tp + tn}{tp+tn+fp+fn}$$

- What fraction of time am I correct in my classification

Precision

$$= \frac{True\ Positives}{My\ Positives} = \frac{tp}{tp+fp}$$

- How much should you trust me when I say that something tests positive
- What fraction of my positives are true positives

Recall = Sensitivity

$$= \frac{True\ Positives}{Real\ Positives} = \frac{tp}{tp+fn}$$

- How much of the reality has been covered by my positive output?
- What fraction of the true positives is captured by my positives? E.g. How many sick people are correctly identified as having the condition?

Specificity

$$= \frac{True\ Negatives}{Real\ Negatives} = \frac{tn}{tn+fp}$$

- How much of the reality has been covered by my negative output?
- What fraction of the true negatives is captured by my negatives? E.g. How many
  identified healthy people do not have the condition?

Real-Life Example: Airport Security Metal Detector

You manage airport security and want to evaluate how well a metal detector finds **dangerous items**.

- **Positive class:** Passenger **has a dangerous item**
- **Negative class:** Passenger **has no dangerous item**

In one day, **10,000 passengers** pass through:

- **50** actually have dangerous items (Positive)
- **9,950** have no dangerous items (Negative)

The detector behaves as follows:

- Correctly flags **40** dangerous passengers → **True Positive (TP = 40)**
- Misses **10** dangerous passengers → **False Negative (FN = 10)**
- Incorrectly flags **100** safe passengers → **False Positive (FP = 100)**
- Correctly clears **9,850** safe passengers → **True Negative (TN = 9,850)**

Confusion Matrix

Total = 10,000 passengers

Predicted ↓ / Actual →	**Dangerous (Pos)**	**Not Dangerous (Neg)**
Predicted Dangerous	**TP = 40**	**FP = 100**
Predicted Safe	**FN = 10**	**TN = 9,850**

Accuracy (overall correctness)

➡ Accuracy = 98.9%
Looks great — but misleading, because the positive class is rare.

Precision (How many flagged dangerous were truly dangerous?)

➡ Precision = 28.6%
Only 28.6% of flagged people were actually dangerous.
Many false alarms!

Recall / Sensitivity (How many dangerous people were caught?)

➡ Recall = 80%
The detector catches most (but not all) dangerous passengers.

Specificity (How many safe people were correctly cleared?)

➡ Specificity = 99%
It is very good at identifying safe passengers.

F1 Score (Balance of Precision & Recall)

➡ F1 Score = 0.42
Moderate performance — reflects poor precision.

What We Actually Want in a Real-Life System
A good security system needs:

High Recall / Sensitivity (VERY IMPORTANT)
We *must* catch nearly **all** dangerous passengers.
Missing even one can be extremely risky.
Ideally:
✔ **Recall > 95%**
✔ **Nearly every dangerous passenger gets flagged**

High Precision

We want fewer false alarms.

Security staff should not be checking hundreds of innocent people.

Ideally:

✔ **Precision > 80%**

✔ When the alarm rings, it is almost always meaningful.

Balanced F1 Score

This shows good harmony between:

- Catching real dangers
- Avoiding false alarms

Ideally:

✔ **F1 > 0.8**

Good Specificity

Letting safe passengers through is also important, but **not as important as recall** in high-risk environments.

Ideally:

✔ **Specificity ~ 98–100%**

Accuracy is the LEAST important

Because the positive class (dangerous passengers) is **rare**.

A worked example

You are shown a set of 21 coins: 10 gold and 11 copper. Your task to accept all gold coins and reject all copper ones

You accept 7 coins as being gold. (these are your positives)

- 5 of these are actually gold
- 2 of these are copper
- You falsely rejected 5 gold ones
- You correctly rejected 9 copper ones

A worked example

You are shown a set of 21 coins: 10 gold and 11 copper. Your task to accept all gold coins and reject all copper ones

You accept 7 coins as being gold. (these are your positives)

- 5 of these are actually gold (these are your true positives, tp)
- 2 of these are copper (these are your false positives, fp)
- You falsely rejected 5 gold ones (false negatives, fn)
- You correctly rejected 9 copper ones (true negatives, tn)

	Actual Gold	**Actual Copper**
Predicted Gold	5	2
Predicted Copper	5	9

- Your **precision** is $\frac{tp}{tp+fp} = \frac{5}{7}$
- Your **recall** is $\frac{tp}{tp+fn} = \frac{5}{10}$
- Your **specificity** is $\frac{tn}{tn+fp} = \frac{9}{11}$
- Your **accuracy** is $\frac{tp+tn}{tp+tn+fp+fn} = \frac{5+9}{5+9+2+5} = \frac{14}{21}$

$$\frac{\text{all of your positives}}{\text{number of actual gold coins}} = \frac{\text{correct answers}}{\text{all attempts}}$$

Realistic extremes

You accept only one coin and that is gold

- Your precision is very high (1/1) but recall is very low (1/10)

	Ac Gld	Ac Cop
Pr Gld	1	0
Pr Cop	9	11

You return all 21 coins

- Your recall is very high (10/10) but precision is very low (10/21)

	Ac Gld	Ac Cop
Pr Gld	10	11
Pr Cop	0	0

Only one out of the 21 coins is gold. And you reject everything.

- Your accuracy is very high (20/21 = 0.95) but precision/recall are 0

	Ac Gld	Ac Cop
Pr Gld	0	0
Pr Cop	1	20

So, what do we do now?

- A combined measure?

		Gold Positive	Gold Negative		
		Gold Labels			
Predicted Labels	**Predicted Positive**	True Positives (tp)	False Positives (fp)	$\frac{tp}{tp + fp}$	"Precision" aka "Positive Predictive Value"
	Predicted Negative	False Negatives (fn)	True Negatives (tn)	$\frac{tn}{fn + tn}$	"Negative Predictive Value"
		$\frac{tp}{tp + fn}$ "Recall" aka "Sensitivity" aka "True Positive Rate"	$\frac{tn}{fp + tn}$ "Specificity" aka "True Negative Rate"		
		$\frac{fn}{tp + fn}$ 1 - Sensitivity = "False Negative Rate" aka "False Rejection Rate"	$\frac{fp}{fp + tn}$ 1 - Specificity = "False Positive Rate" aka "False Acceptance Rate"		

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn}$$

True Positive Rate = "Positive Likelihood Ratio" False Positive Rate

False Negative Rate = "Negative Likelihood Ratio"
True Negative Rate

Probability = "Odds," often expressed as X:Y 1 - Probability

Evaluation of Classifiers

The F-Score

Agha Ali Raza

CS535/EE514 – Machine

		Gold Positive	Gold Negative		
		Gold Labels			
Predicted Labels	**Predicted Positive**	True Positives (tp)	False Positives (fp)	$\frac{tp}{tp + fp}$	"Precision" aka "Positive Predictive Value"
	Predicted Negative	False Negatives (fn)	True Negatives (tn)	$\frac{tn}{fn + tn}$	"Negative Predictive Value"
		$\frac{tp}{tp + fn}$ "Recall" aka "Sensitivity" aka "True Positive Rate"	$\frac{tn}{fp + tn}$ "Specificity" aka "True Negative Rate"		
		$\frac{fn}{tp + fn}$ 1 - Sensitivity = "False Negative Rate" aka "False Rejection Rate"	$\frac{fp}{fp + tn}$ 1 - Specificity = "False Positive Rate" aka "False Acceptance Rate"		

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn}$$

True Positive Rate = "Positive Likelihood Ratio" False Positive Rate

False Negative Rate = "Negative Likelihood Ratio"
True Negative Rate

Probability = "Odds," often expressed as X:Y 1 - Probability

A worked example

You are shown a set of 21 coins: 10 gold and 11 copper. Your task to accept all gold coins and reject all copper ones

You accept 7 coins as being gold (these are your positives)

- 5 of these are actually gold (these are your true positives, tp)
- 2 of these are copper (these are your false positives, fp)
- You falsely rejected 5 gold ones (false negatives, fn)
- You correctly rejected 9 copper ones (true negatives, tn)

	Actual Gold	**Actual Copper**
Predicted Gold	5	2
Predicted Copper	5	9

- Your **precision** is $\frac{tp}{tp+fp} = \frac{5}{7}$
- Your **recall** is $\frac{tp}{tp+fn} = \frac{5}{10}$
- Your **specificity** is $\frac{tn}{tn+fp} = \frac{9}{11}$
- Your **accuracy** is $\frac{tp+tn}{tp+tn+fp+fn} = \frac{5+9}{5+9+2+5} = \frac{14}{21}$

$$\frac{\text{true positives}}{\text{all of your positives}} = \frac{tp}{tp+fp} = \frac{5}{7}$$

$$\frac{\text{true positives}}{\text{number of actual gold coins}} = \frac{tp}{tp+fn} = \frac{5}{10}$$

$$\frac{\text{true negatives}}{\text{number of copper coins}} = \frac{tn}{tn+fp} = \frac{9}{11}$$

$$\frac{\text{correct answers}}{\text{all attempts}} = \frac{tp+tn}{tp+tn+fp+fn} = \frac{5+9}{5+9+2+5} = \frac{14}{21}$$

Realistic extremes

You accept only one coin and that is gold

- Your precision is very high (1/1) but recall is very low (1/10)

	Ac Gld	Ac Cop
Pr Gld	1	0
Pr Cop	9	11

You return all 21 coins

- Your recall is very high (10/10) but precision is very low (10/21)

	Ac Gld	Ac Cop
Pr Gld	10	11
Pr Cop	0	0

Only one out of the 21 coins is gold. And you reject everything.

- Your accuracy is very high (20/21 = 0.95) but precision/recall are 0

	Ac Gld	Ac Cop
Pr Gld	0	0
Pr Cop	1	20

So, what do we do now?

- **A combined measure?**

How to Combine Precision and Recall (F1-Score)?

- The ratio: $\frac{Precision}{Recall}$ may work but does not have a way of differentially weighting Precision and Recall
- How about we take a mean of the two?
- But what kind of a mean should we use?
- Simple Arithmetic Mean is problematic: e.g.
  - P = 0.0, R = 1.0, AM = 0.5
  - P = 0.1, R = 0.9, AM = 0.5
  - P = 0.5, R = 0.5, AM = 0.5
- Requirements:
  - We need a weighted mean as we may care more about P or R
  - We need a conservative (deliberately lower) estimate of mean
  - If P and R are far apart, we need the mean to tend to the lower value
  - In order to do well, a classifier must do well on both P and R so that it cannot beat the system by being either too selective or too indiscriminate

Types of Means

Arithmetic Mean

$$AM = \frac{a_1 + a_2 + a_3 + \cdots + a_n}{n}$$

For 2 values:

$$AM = \frac{a_1 + a_2}{2}$$

Geometric Mean

$$GM = \sqrt[n]{a_1 . a_2 . a_3}$$

For 2 values:

$$GM = \sqrt{a_1 a_2}$$

Harmonic Mean

$$HM = \frac{1}{\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} + \cdots + \frac{1}{a_n}}$$

For 2 values:

$$HM = \frac{2}{\frac{1}{a_1} + \frac{1}{a_2}} = \frac{2 a_1 a_2}{a_1 + a_2}$$

Harmonic Mean

Ref:
http://economistatlarge.com/finance/applied-finance/differences-arithmetic-geometric-harmonic-means

- The harmonic mean of a set of numbers is the reciprocal of the arithmetic mean of reciprocals
- Best used in situations where extreme outliers exist in the population

x0	x1	x2	x3	x4	x5	x6	x7	x8	AM	GM	HM
1	2	3	4	5	6	7	8	9	5.00	4.15	3.18
2	4	8	16	32	64	128	256	512	113.56	32.00	9.02
5	5	5	5	5	5	5	5	5	5.00	5.00	5.00
5	5	5	5	5	5	5	5	**10**	5.56	5.40	5.29
5	5	5	5	5	5	5	5	**100**	15.56	6.97	5.59
5	5	5	5	5	5	5	5	**1000**	115.56	9.01	5.62
5	5	5	5	5	5	5	5	**10000**	1115.56	11.63	5.62
5	5	5	5	5	5	5	5	**100000**	11115.56	15.03	5.62
5	5	5	5	5	5	5	100000	**100000**	22226.11	45.16	6.43
5	5	5	5	5	100000	100000	100000	**100000**	44447.22	407.89	9.00
5	100000	100000	100000	100000	100000	100000	100000	**100000**	88889.44	33274.21	44.98
100000	100000	100000	100000	100000	100000	100000	100000	**100000**	100000.0	100000.0	100000.0

F-1-MEASURE

- The harmonic mean of P and R:
  - Is high when both P and R are high.
  - Is low when even one of P and R is low.
- A combined measure that assesses the P/R tradeoff is F measure (weighted harmonic mean of precision and recall):

$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P + R}$$

F-MEASURE

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

The β parameter differentially weights the importance of recall and precision

- based on the needs of an application
- values of $\beta > 1$ favor recall, while
- values of $\beta < 1$ favor precision

When $\beta = 1$, precision and recall are equally balanced

This is the most frequently used metric, and is called F_β=1 or just $F1$:

$$F_1 = \frac{2PR}{P + R}$$

F-β-MEASURE

We can choose to favor precision or recall by using an interpolation weight α:

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} \qquad F = \frac{(\beta^2+1)PR}{\beta^2 P + R} \qquad \beta = \frac{1-\alpha}{\alpha}$$

$$F = \frac{1}{\left(\frac{1}{\beta^2+1}\frac{1}{P} + \frac{\beta^2}{\beta^2+1}\frac{1}{R}\right)}$$

$$F_\beta = \frac{1}{\frac{1}{\beta+1}\frac{1}{\text{precision}} + \frac{\beta}{\beta+1}\frac{1}{\text{recall}}} = (1+\beta)\frac{\text{precision} \cdot \text{recall}}{\beta \cdot \text{precision} + \text{recall}}$$

- Balanced F1 measure has $\beta = 1$ (that is, $\alpha = ½$) as shown above.
- To give more weight to the Precision, we pick a β value in the interval
  - $0 < \beta < 1$.
  - notice that it is getting multiplied with P in the denominator
- To give more weight to the Recall, we pick a β Value in the interval
  - $1 < \beta < +\infty$

F-β-MEASURE

We can also set to favor precision or recall by using an interpolation weight α:

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

$$= \frac{(\beta^2+1)PR}{\beta^2 P + R} \qquad \beta = \frac{1-\alpha}{\alpha}$$

$$F = \frac{1}{\left(\frac{1}{\beta^2+1}\frac{1}{P} + \frac{\beta^2}{\beta^2+1}\frac{1}{R}\right)}$$

$$F_\beta = \frac{1}{\frac{1}{\beta+1}\frac{1}{\text{precision}} + \frac{\beta}{\beta+1}\frac{1}{\text{recall}}}$$

$$= (1+\beta)\frac{\text{precision}\cdot\text{recall}}{\beta\cdot\text{precision}+\text{recall}}$$

- $\beta \to 0$ considers only precision

$$\frac{(\beta^2+1)PR}{\beta^2 P + R} = \frac{\beta^2 PR + PR}{\beta^2 P + R} = \frac{PR}{R} = P$$

- $\beta \to \infty$ only recall

$$\frac{1}{\frac{1}{\beta^2+1}\frac{1}{P} + \frac{\beta^2}{\beta^2+1}\frac{1}{R}} = \frac{1}{0\frac{1}{P} + 1\frac{1}{R}} = R$$

$\lim_{n\to\infty} \frac{n}{n+1} \to 1$ as $\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{5}{6}, \frac{6}{7}, \frac{7}{8}$ ... gets closer to 1

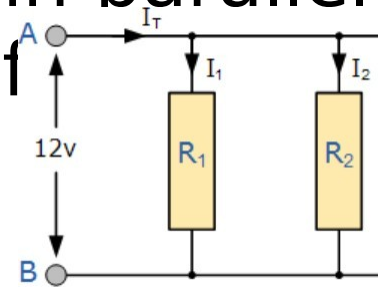
F-1, F-0.5,

Precision	Recall	F-1	F-0.5	F-2
0	1	0	0	0
0.1	0.9	0.18	0.121622	0.346154
0.2	0.8	0.32	0.235294	0.5
0.3	0.7	0.42	0.33871	0.552632
0.4	0.6	0.48	0.428571	0.545455
0.5	0.5	0.5	0.5	0.5
0.6	0.4	0.48	0.545455	0.428571
0.7	0.3	0.42	0.552632	0.33871
0.8	0.2	0.32	0.5	0.235294
0.9	0.1	0.18	0.346154	0.121622
1	0	0	0	0
1	1	1	1	1
0.5	1	0.666667	0.555556	0.833333
1	0.5	0.666667	0.833333	0.555556
0.1	1	0.181818	0.121951	0.357143
1	0.1	0.181818	0.357143	0.121951

Intuition

Ref: https://stats.stackexchange.com/questions/49226/how-to-interpret-f-measure-values, http://www.electronics-tutorials.ws/resistor/res_4.html

The formula for F-measure (F1, with β=1) is the same as the formula giving the equivalent resistance composed of two resistances placed in parallel in physics (forgetting about the f



$$\frac{1}{R_T} = \frac{1}{R_1} + \frac{1}{R_2} \qquad R_T = \frac{R_1 \times R_2}{R_1 + R_2}$$

This analogy would define F-measure as the equivalent resistance formed by *sensitivity* and *precision* placed in parallel.

Net resistance gets reduced as soon as any one among the two looses resistance

Evaluation of Classifiers

More than 2 Classes

Source

- Jurafsky and Martin, SLP3, https://web.stanford.edu/~jurafsky/slp3/

More than two classes

Many classification tasks in language processing have more than two classes:

- Sentiment analysis (positive, negative, neutral),
- Part-of-speech tagging (|POS tags|)
- Word sense disambiguation (|word senses|)
- Emotion detection (|emotions|)

More Than Two Classes: Sets of binary classifiers

Any-of or multi-label classification

- An instance can belong to 0, 1, or more than one classes.

For each class $c \in C$

- Build a binary classifier h_c to distinguish c from all other classes
  $c' \in C$ (c vs not c)

Given test instance d,

- Evaluate it for membership in each class using each h_c
- Each h_c creates a score s_c , and returns 1 if s_c is greater than the defined threshold
- d belongs to any class for which h_c returns true

More Than Two Classes: Sets of binary classifiers

One-of or multinomial classification

- Classes are mutually exclusive: each instance in exactly one class

For each class $c \in C$

- Build a classifier h_c to distinguish c from all other classes $c' \in C$.

Given test instance d,

- Evaluate it for membership in each class using each h_c
- d belongs to the one class with maximum score

Evaluation

- Say ***one-of*** email categorization decision (urgent, normal, spam)

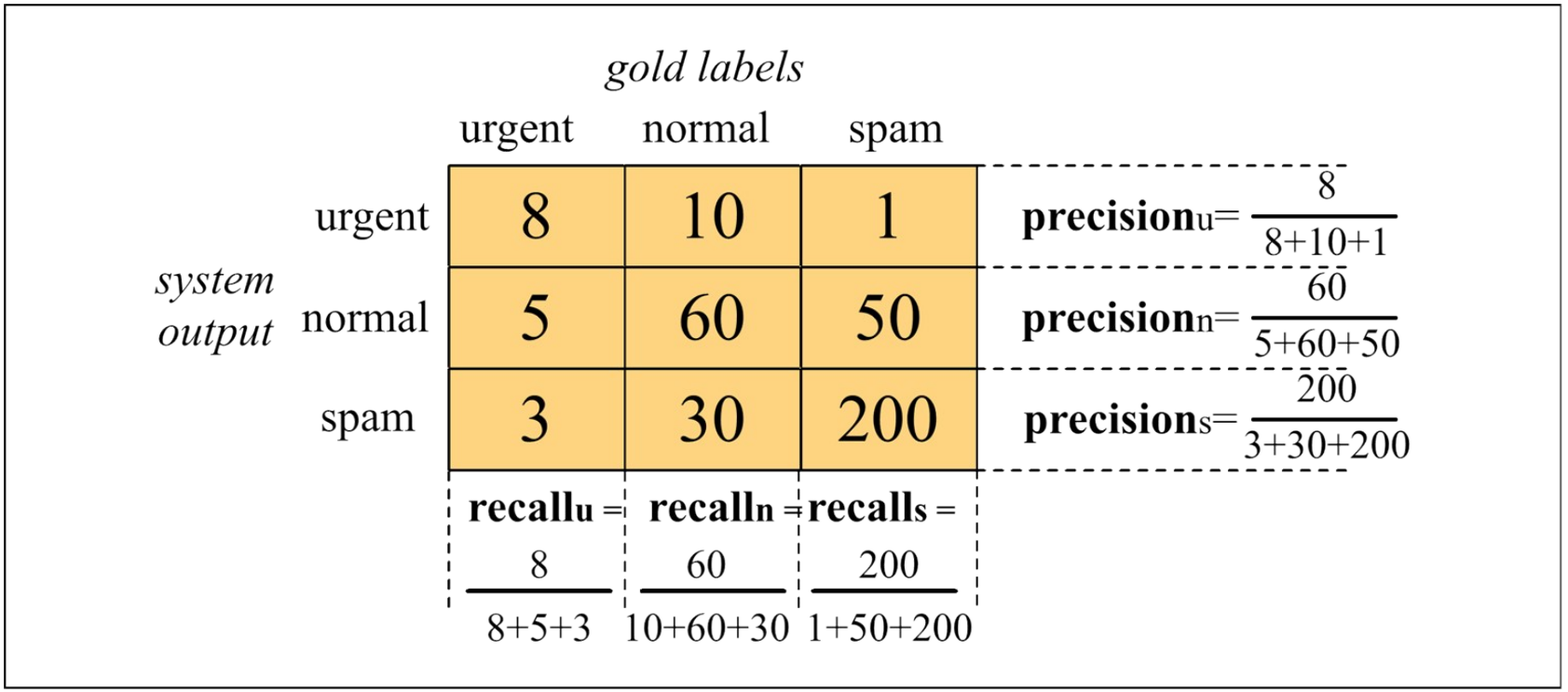
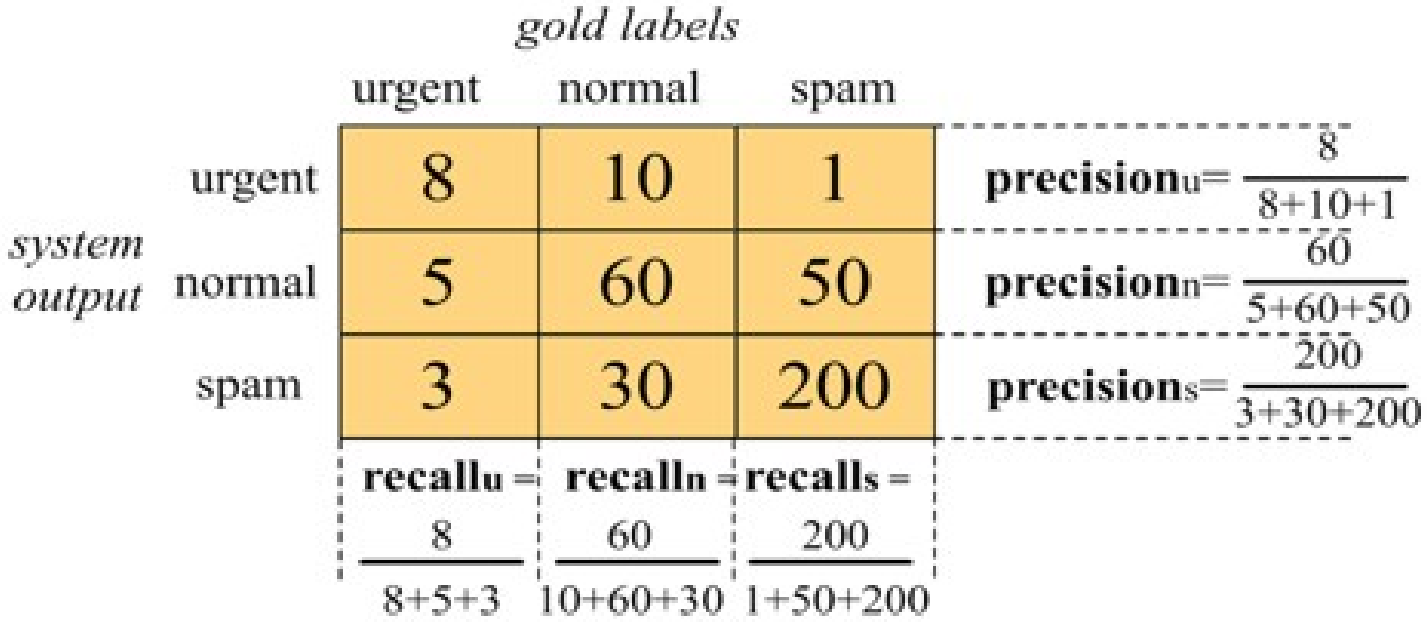


Figure 6.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2), how many documents from c_1 were (in)correctly assigned to c_2



Meaning examples:

- **8** → Urgent emails correctly predicted as urgent
- **10** → Normal emails wrongly predicted as urgent
- **1** → Spam predicted as urgent
- **60** → Normal correctly predicted as normal
- **50** → Spam wrongly predicted as normal
- **200** → Spam correctly predicted as spam

Micro- vs. Macro-Averaging

If we have more than one class, how do we combine multiple performance measures into one quantity?

1. **Macro-averaging**: Compute performance for each class, then average.
   Steps:

   Convert each class into **binary (One-vs-Rest)**.
   Compute **precision / recall / F1** for that class.
   Average the values.

2. **Micro-averaging**: Collect decisions for all classes, compute contingency table, evaluate
   Steps: Sum **True Positives of all classes**
   Sum **False Positives of all classes**
   Sum **False Negatives of all classes**
   Then compute precision/recall.

Evaluation

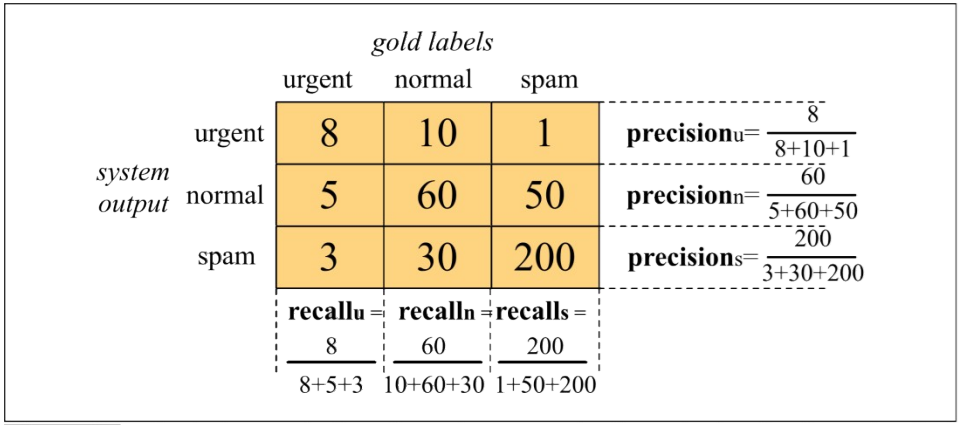


Figure 6.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2), how many documents from c_1 were (in)correctly assigned to c_2

Macro Averaging

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

Micro Averaging

Figure 6.6 Separate contingency tables for the 3 classes from the previous figure, showing the pooled contingency table and the microaveraged and macroaveraged precision.

Evaluation

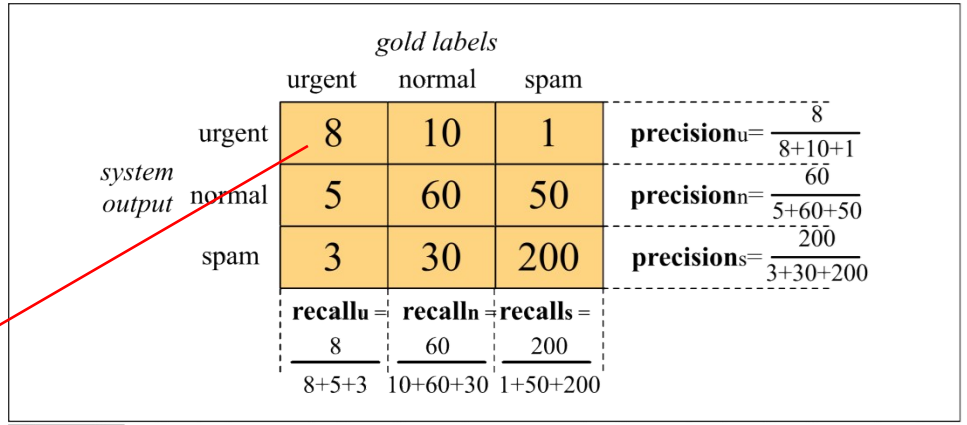


Figure 6.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2), how many documents from c_1 were (in)correctly assigned to c_2

Macro Averaging

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

Micro Averaging

Figure 6.6 Separate contingency tables for the 3 classes from the previous figure, showing the pooled contingency table and the microaveraged and macroaveraged precision.

Evaluation

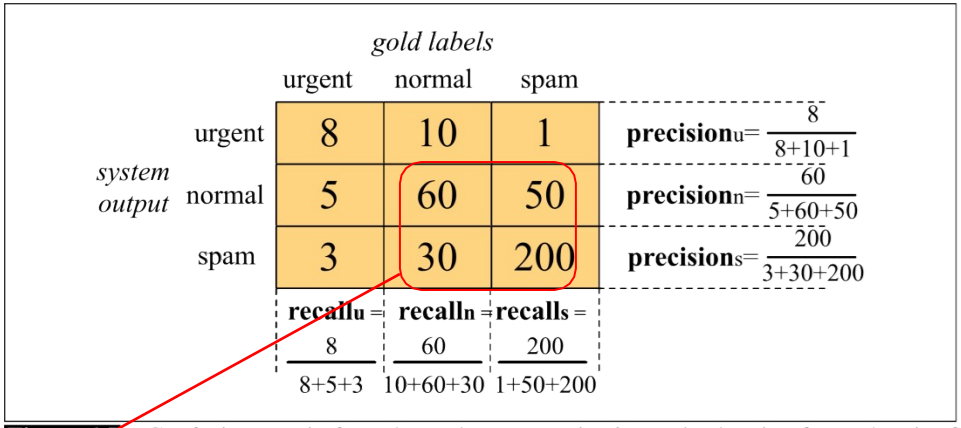


Figure 6.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2), how many documents from c_1 were (in)correctly assigned to c_2

Macro Averaging

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

Micro Averaging

Figure 6.6 Separate contingency tables for the 3 classes from the previous figure, showing the pooled contingency table and the microaveraged and macroaveraged precision.

Evaluation

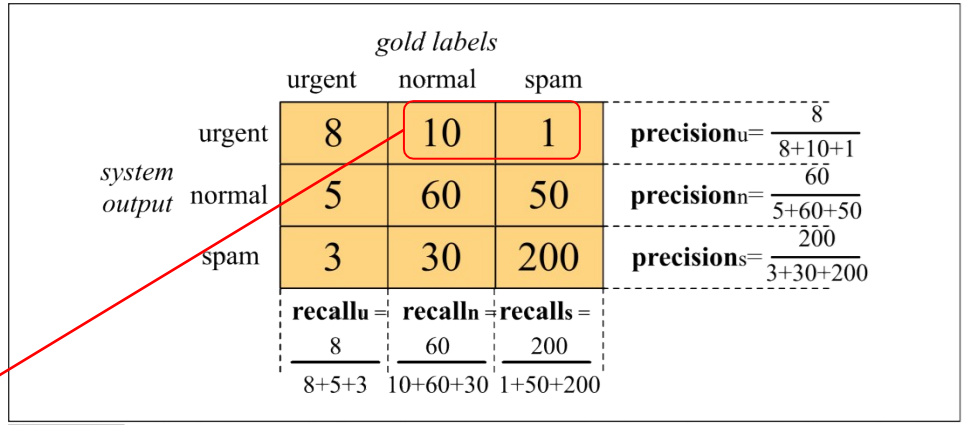


Figure 6.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2), how many documents from c_1 were (in)correctly assigned to c_2

Macro Averaging

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

Micro Averaging

Figure 6.6 Separate contingency tables for the 3 classes from the previous figure, showing the pooled contingency table and the microaveraged and macroaveraged precision.

Evaluation

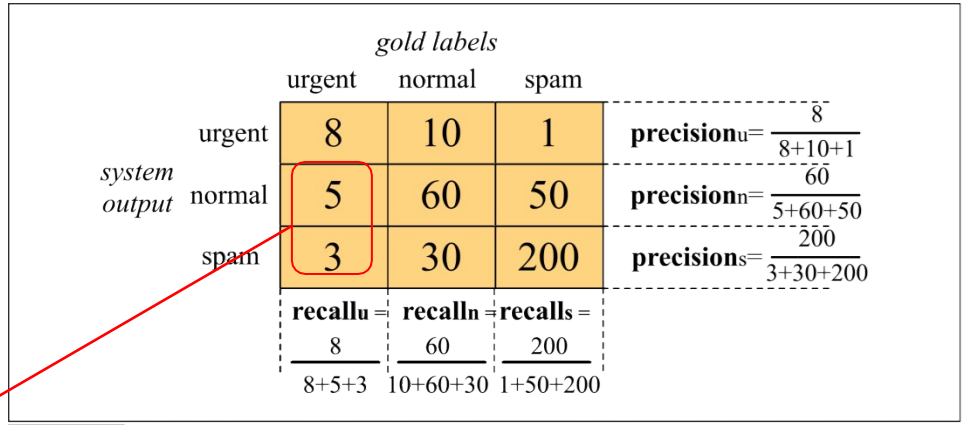


Figure 6.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2), how many documents from c_1 were (in)correctly assigned to c_2

Macro Averaging

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

Micro Averaging

Figure 6.6 Separate contingency tables for the 3 classes from the previous figure, showing the pooled contingency table and the microaveraged and macroaveraged precision.

Evaluation

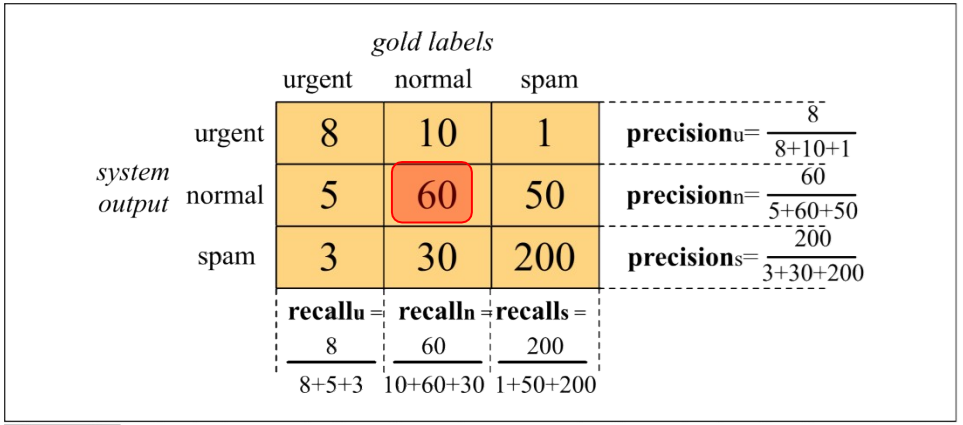


Figure 6.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2), how many documents from c_1 were (in)correctly assigned to c_2

Macro Averaging

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

Micro Averaging

Figure 6.6 Separate contingency tables for the 3 classes from the previous figure, showing the pooled contingency table and the microaveraged and macroaveraged precision.

Evaluation

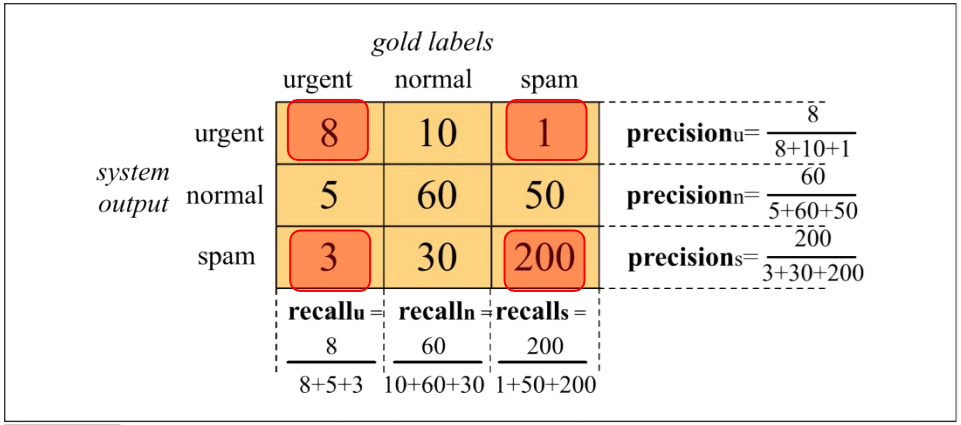


Figure 6.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2), how many documents from c_1 were (in)correctly assigned to c_2

Macro Averaging

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

Micro Averaging

Figure 6.6 Separate contingency tables for the 3 classes from the previous figure, showing the pooled contingency table and the microaveraged and macroaveraged precision.

Evaluation

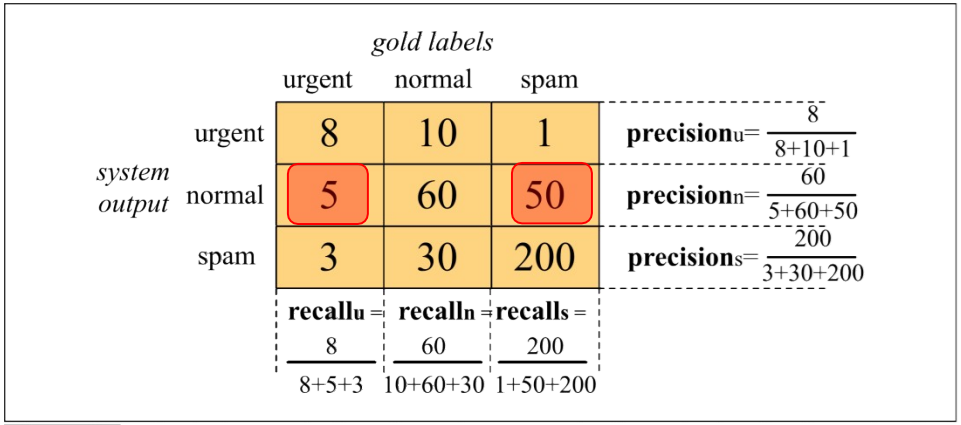


Figure 6.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2), how many documents from c_1 were (in)correctly assigned to c_2

Macro Averaging

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

Micro Averaging

Figure 6.6 Separate contingency tables for the 3 classes from the previous figure, showing the pooled contingency table and the microaveraged and macroaveraged precision.

Evaluation

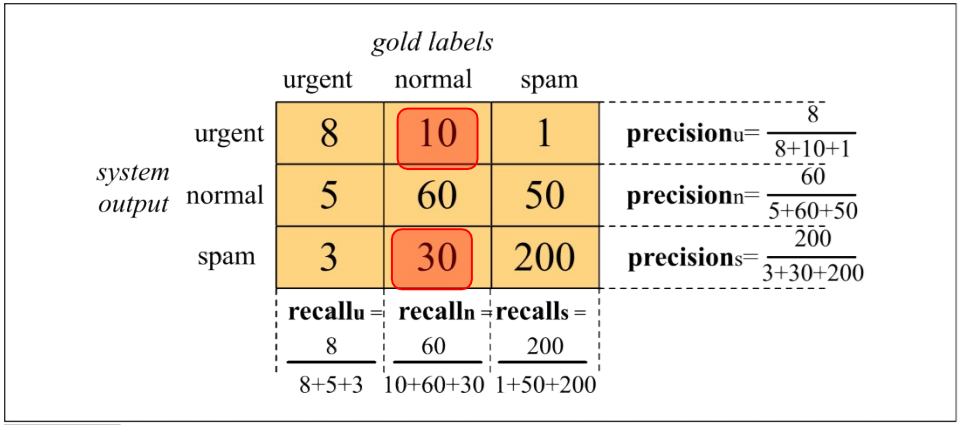


Figure 6.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2), how many documents from c_1 were (in)correctly assigned to c_2

Macro Averaging

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

Micro Averaging

Figure 6.6 Separate contingency tables for the 3 classes from the previous figure, showing the pooled contingency table and the microaveraged and macroaveraged precision.

Evaluation

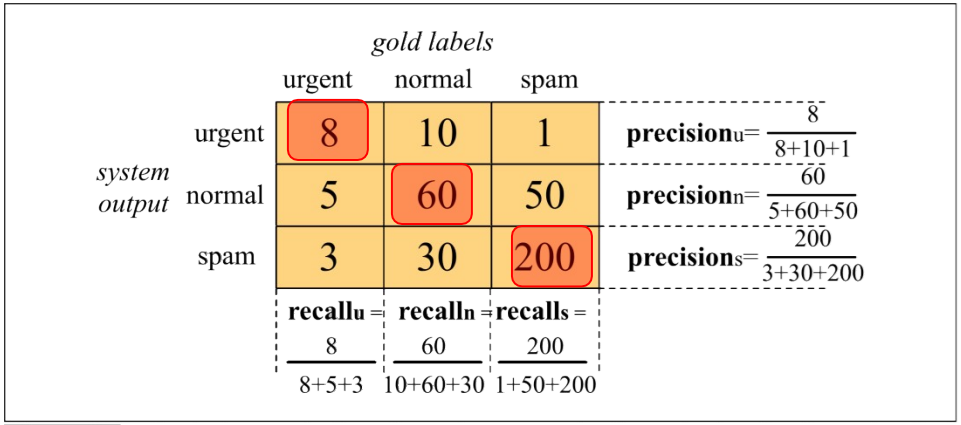


Figure 6.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2), how many documents from c_1 were (in)correctly assigned to c_2

Macro Averaging

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

Micro Averaging

Figure 6.6 Separate contingency tables for the 3 classes from the previous figure, showing the pooled contingency table and the microaveraged and macroaveraged precision.

Evaluation

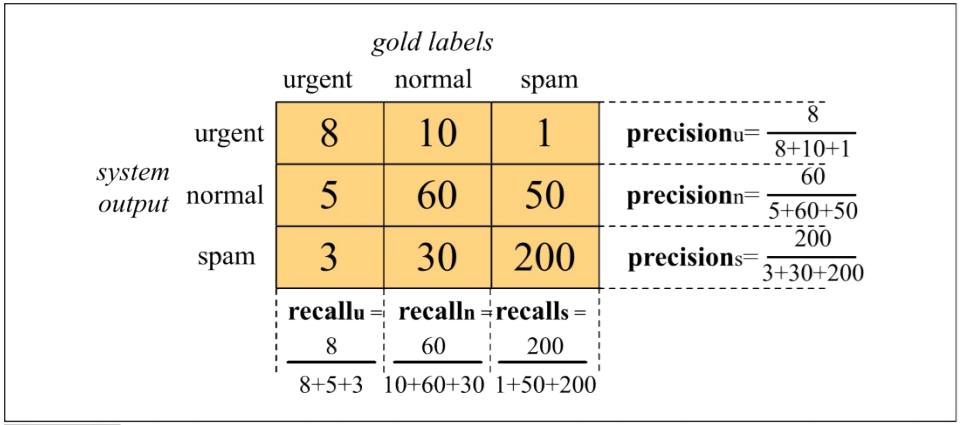


Figure 6.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2), how many documents from c_1 were (in)correctly assigned to c_2

Macro Averaging

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

Micro Averaging

Figure 6.6 Separate contingency tables for the 3 classes from the previous figure, showing the pooled contingency table and the microaveraged and macroaveraged precision.

Evaluation

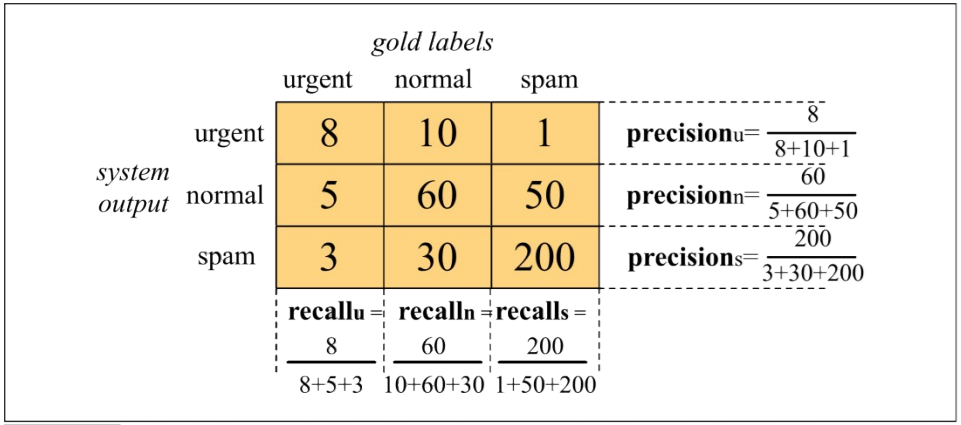


Figure 6.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2), how many documents from c_1 were (in)correctly assigned to c_2

Macro Averaging

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

Micro Averaging

Figure 6.6 Separate contingency tables for the 3 classes from the previous figure, showing the pooled contingency table and the microaveraged and macroaveraged precision.

Evaluation

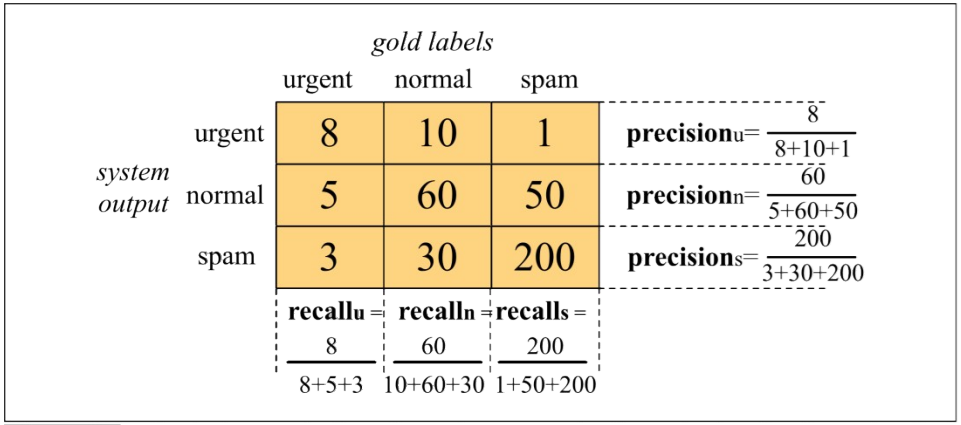


Figure 6.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2), how many documents from c_1 were (in)correctly assigned to c_2

Macro Averaging

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

Micro Averaging

Figure 6.6 Separate contingency tables for the 3 classes from the previous figure, showing the pooled contingency table and the microaveraged and macroaveraged precision.

Evaluation

- A micro-average is dominated by the more frequent class (in this case spam)
  - as the counts are pooled
- The macro-average better reflects the statistics of the smaller classes,
  - is more appropriate when performance on all the classes is equally important.

Evaluation of Classifiers

The Gold Labels

Agha Ali Raza

The Output of a

#	Height (inches)	Weight (kgs)	B.P. Sys →	B.P. Dia	Heart disease →	
	𝑋				𝑦	𝑦(𝑋)
1	62	70	120	80	**No**	No
2	72	90	110	70	**No**	Yes
3	74	80	130	70	**No**	No
4	65	120	150	90	**Yes**	Yes
5	67	100	140	85	**Yes**	No
6	64	110	130	90	**No**	Yes
7	69	150	170	100	**Yes**	Yes
8	75	127	160	95	**Yes**	No
9	66	66	135	90	**Yes**	Yes

- y: Gold standards / Gold labels / Ground truth
- $y(X)$: Predicted labels

(How) can we trust the Gold Labels?

- Objective data gathered from the real world:
  - Measurement of heights, weights, weather phenomena
    - Aberrations, sensor malfunctions and errors
- Subjective data, annotated by humans (experts or community)
  - Tagging emotions, SPAM/HAM, sentiment, misinformation
    - Errors of judgement, biases, human error
- Use multiple sources for each label
  - Multiple sensors measuring the same phenomenon
  - Multiple humans annotating the same data
    - Inter-annotator agreements
  - Resource intense
    - Partial overlap of data
- Is annotator agreement the answer to all our worries?
  - Distribution of annotators
  - Chance agreements

Intuition of Agreement (and Randomness)

- Annotator1 (A1) and **Annotator2 (A2)** are deciding if posts should be deleted from a social media platform.
- Most of the posts should not be deleted.
- A1 is not an expert and just randomly selects 1% of posts.
- A2 is an expert and very carefully selects 1% of posts.
- A1 and A2 still agree on at least 98% of posts (that need not be deleted). Yaayyy!! ??
- So, what is the problem?
- The problem is that A1 was just picking posts at random. Because most posts are not deleted, A1 and A2 agree most of the time.
- But that does not mean that A1 and A2 are using similar ratings.
- Cohen's Kappa corrects for this – It would be approximately zero in this case.

https://www.quora.com/What-is-an-intuitive-explanation-of-Cohens-kappa-statistic

Gold Labels, Annotators and Agreement

Multiple

Utterances	Ann1	Ann2	Raw Agreement	Ann Rand	Agreement (A1, Rand)	Agreement (A2, Rand)
S1	+	+	1	-	0	0
S2	-	-	1	+	0	0
S3	+	-	0	+	1	0
S4	-	+	0	-	1	0
…	…	…	…	…	…	…
			Sum of Agreements		Sum of chance agreements	Sum of chance agreements

- Chance agreements
- *Cohen's Kappa* ($\kappa = \frac{P_o - P_e}{1 - P_e} = 1 - \frac{1 - P_o}{1 - P_e}$) and *Krippendorff's alpha* ($\alpha = 1 - \frac{D_o}{D_e}$, where D is disagreement)
- Agreement over a random subset with an expert annotator

https://towardsdatascience.com/cohens-kappa-9786ceceab58#:~:text=Cohen's%20kappa%20measures%20the%20agreement,raters%20may%20agree%20by%20chance

.

https://repository.upenn.edu/cgi/viewcontent.cgi?article=1043&context=asc_papers#:~:text=Krippendorff's%20alpha%20(%CE%B1)%20is%20a,assign%20computable%20values%20to%20them.

Cohen's Kappa

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

- P_o is the probability of actual agreement
- P_e is the expected probability of "agreement by chance"- the probability of the agreement if the rating is random
- The numerator is the difference between the observed agreement and chance agreement
- The denominator is the difference between the maximum possible observed agreement (i.e., 1: 100% agreement) and the chance agreement
- As the denominator is just a normalizer, observe that κ is high when the difference between the observed and chance agreement is the highest
  - For a good model, the observed difference and the maximum difference are close to each other, and Cohen's kappa is close to 1.
  - For a random model, the overall accuracy is all due to random chance, the numerator is 0, and Cohen's kappa is 0.
  - Cohen's kappa could also theoretically be negative. Then, the overall accuracy of the model would be even lower than what could have been obtained by a random guess.

Kappa value interpretation Landis & Koch (1977):

<0 No agreement

0 — .20 Slight

.21 — .40 Fair

.41 — .60 Moderate

.61 — .80 Substantial

.81–1.0 Perfect

https://thenewstack.io/cohens-kappa-what-it-is-when-to-use-it-and-how-to-avoid-its-pitfalls/

Cohen's Kappa

Suppose that you were analyzing data related to a group of 50 people applying for a grant. Each grant proposal was read by two readers and each reader either said "Yes" or "No" to the proposal. Suppose the disagreement count data were as follows, where A and

		B	
		Yes	No
A	Yes	a	b
	No	c	d

		B	
		Yes	No
A	Yes	20	5
	No	10	15

- The observed proportionate agreement is:

$$p_o = \frac{a+d}{a+b+c+d} = \frac{20+15}{50} = 0.7$$

- To calculate p_e (the probability of random agreement) we note that:
  - Reader A said "Yes" to 25 applicants and "No" to 25 applicants. Thus reader A said "Yes" 50% of the time.
  - Reader B said "Yes" to 30 applicants and "No" to 20 applicants. Thus reader B said "Yes" 60% of the time.
- So the expected probabil

$$p_{\text{Yes}} = \frac{a+b}{a+b+c+d} \cdot \frac{a+c}{a+b+c+d} = 0.5 \times 0.6 = 0.3$$

- Similarl y,

$$p_{\text{No}} = \frac{c+d}{a+b+c+d} \cdot \frac{b+d}{a+b+c+d} = 0.5 \times 0.4 = 0.2$$

- Overall random agreement probability is the probability that they agreed on either Yes or No, i.e.:

$$p_e = p_{\text{Yes}} + p_{\text{No}} = 0.3 + 0.2 = 0.5$$

- So now applying our formula for Cohen's Kappa we get:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

Kappa value interpretation Landis & Koch (1977):

<0 No agreement

0 — .20 Slight

.21 — .40 Fair

.41 — .60 Moderate

.61 — .80 Substantial

.81–1.0 Perfect

From: https://en.wikipedia.org/wiki/Cohen%27s_kappa